Exhaustive Enumeration of Molecular Substructures

RICHARD G. A. BONE and HUGO O. VILLAR*

Terrapin Technologies, Inc., 750-H Gateway Boulevard, South San Francisco, California 94080-7020

Received 28 November 1995; accepted 8 April 1996

ABSTRACT

This article addresses the systematic and complete enumeration of all the substructures of any size present in a given molecule. The study is not restricted to features which could be defined *a priori* such as rings or chains. Contrary to prior expectation the exhaustive enumeration is tractable with current computational tools. Results are presented for several families of skeletons which are widespread in chemistry. It is shown that the numbers of constituent substructures of each size are related to the molecular topology, in particular the degree of branching. The number substructures which are distinct depends additionally on the number of different atom and bond types present. The overall shapes of the distribution of substructure counts as a function of substructure size are found to be similar within particular classes of molecules. These distributions are compared and found to be characteristic of certain topologies. For several simple classes of molecule, analytic expressions are provided for the numbers of substructures as a function of fragment and molecule size. These results hold promise for identifying potentially useful scaffolds for use in combinatorial chemistry.

Introduction

The analysis of a molecule's constituent fragments has importance throughout chemistry[1] and has already received much attention in the realms of, for example, organic synthesis planning (''retrosynthetic analysis''),[2] NMR spectra prediction,[3] calculation of partition coefficients,[4,5] and elucidation of fragmentation mechanisms in mass spectrometry.[6] In biological applications, analysis of molecular fragments has proven extremely useful in, for example, the computer-aided evaluation of chemical toxicity[7,8] and in the search for analogs in chemical databases.[9]

Substructural analysis has recently been given further impetus by the expansion of combinatorial methods in medicinal chemistry.[10,11] These power-

This article includes Supplementary Material available from the authors upon request or via the Internet at ftp.wiley.com/public/journals/jcc/suppmat/18/86 or http://www.wiley.com/jcc

*Author to whom all correspondence should be addressed. E-mail: Hugo_Villar@trpntech.com

CCC 0192-8651 / 97 / 010086-22

ful techniques are based on the automated generation of combinations of chosen molecular building blocks and they have led to the rapid creation of libraries containing very large numbers of new molecules.[12] To increase the chances of finding a biologically active molecule in a given library, a diverse selection of molecules should be generated. Molecular recognition depends on the presence of key substructures and molecular properties are often influenced heavily by the underlying substructural composition. Therefore, it is desirable to use building blocks which give rise to large numbers of substructures with few repetitions.[13] Such an approach effectively maximizes the number of distinct substructures inherent in a library generated in this way.

Empirical[14] and theoretical[13] methods for quantifying diversity are currently active areas of research. Definitions of what constitutes ''molecular diversity,'' the identification of ''useful'' molecular fragments, and the ways in which they may be assembled into novel structures are of much interest.[15, 16] The need to ensure diversity as well as to identify the types of substructures that achieve diversity most effectively, stresses the importance of analyzing the substructural content of commonly used molecules. A first step toward rationalizing this process is to count the numbers and types of substructural features present in a molecule or set of molecules. Such a study involves the computation of all possible fragments of each size as well as the number of times that each distinct fragment appears.

It has been a consistent expectation that the numbers of possible fragments in a molecule would increase dramatically with their size and that their enumeration would rapidly become prohibitive. This presumed ''combinatorial explosion'' has deterred previous efforts aimed at the generation of all possible fragments as well as the examination of the counts of large fragments, in all but the simplest of cases. Therefore, to make the enumeration of substructures tractable previous approaches have input the chemist's viewpoint at a very early stage so that the fragments considered are based on rational mechanistic understandings of how molecules may be constructed or broken down.[2, 6]

This article analyses, at the two-dimensional level, the total number of fragments that can be cut out of a molecular skeleton as a function of their size. This necessitates, not just the generation of all the possible substructures from a given skeleton, but the identification of duplicates via a thorough comparison of those substructures with one another.

The feasibility of such an analysis is illustrated by the study of several homologous series of simple molecules and analogs in which heteroatom substitution is included in systematic ways. Such an exercise illuminates the link between the form of a molecular skeleton and the numbers of building blocks it comprises. Explanations can be provided, in each case, for the observed dependence of substructure counts on molecular size and topology. In this way the influence of common chemical features on the distributions of fragments of all possible sizes can be clearly identified. The dominant influences on substructure enumerations will be shown to be atom connectivities in the skeleton and heteroatom usage. This systematic analysis of characteristic topologies facilitates an understanding of which types of core structures most appropriately satisfy the criteria for molecular diversity.

Methodology

The description of a molecule in a two-dimensional (2D) chemical database is sufficient for an enumeration of all its substructures. The standard representation of a molecule at the 2D level is as a list of atom types and a connection table, that is, a list of pairs of atoms bonded to one another with a specified bond type.

As is normally done for 2D databases, hydrogens are omitted; restricting the analysis to ''heavy atoms'' makes the problem more manageable because the number of fragments generated depends on the number of atoms in the molecule under consideration. Even so, this restriction does not result in loss of information because the knowledge of bond types between the heavy atoms and standard assumptions of their respective valencies allows regeneration of full structural formulae for both the parent molecule and its fragments.

From the connection table it is trivial to construct the molecular graph, and techniques of graph theory[17] may then be applied. At this point a few definitions are introduced.

The N vertices (nodes) of the graph represent the atoms and are identified by numerical labels $(1, 2, \ldots, N)$, and their atom types (C, O, N, S, etc.). Note the technical distinction between the atom ''label'' (which derives from an arbitrarily chosen numbering scheme) and the atom ''type'' which defines the chemistry of the problem. The

''size'' of a molecule may be quantified by the number N itself. The E edges of the graph represent the bonds between pairs of atoms and carry integer ''weights'' corresponding with chemists' conventional understanding of bond order. Hydrogen bonds are not considered. The ''connectivity,'' d, of a given atom is the number of other atoms to which it is bonded, irrespective of bond type. The graph is said to be ''connected'' if it is possible to trace paths along consecutive edges from every vertex to every other. All the graphs dealt with here are connected.

The first step in the analysis is to generate all connected subgraphs of the ''parent'' molecular graph which contain distinct sets of vertices by label. A connected subgraph corresponds to a molecular fragment (which need not be capable of independent existence in its own right); at this stage, it is uniquely defined by the list of atom labels which comprise it. Therefore, two subgraphs of the same size are distinct if they differ by at least one vertex label. For a given molecule with N atoms, the number of distinct subgraphs of size m will be represented by the distribution function, $S(N, m)$, where m ranges from 1 (single atoms) to N (the parent molecule).

By way of example, for the skeleton methylcyclopentane, all the distinct subgraphs are listed by vertex label in Figure 1a. At the bottom of the figure are the totals for each size. Note that, at this stage in the analysis, the number of distinct subgraphs of each size that are found is independent of the atom types and bond types and, of course, the way in which the atoms are labeled. Thus, in Figure 1a, 2-methyltetrahydrofuran, 3-methyltetrahydrofuran, and hydroxycyclopentane would give identical distributions of subgraphs. In this context, such groups of molecules are referred to as ''isoskeletal.''

It should be noted that, in this study, any pair of atoms bonded to one another in the parent molecule retain that bond in any fragment which contains them both.[18] Therefore, there is only one five-atom subgraph which contains all five ring atoms; that is, ''chains'' of all five ring atoms are not found. Such a result is a consequence of defining subgraphs by sets of vertex-labels rather than sets of edge-labels which would be an equally valid, but distinct, approach. The rationale behind this choice is that a ''substructure'' is envisaged to be a set of atoms and the bonds between them which is extracted from the parent molecule, *in toto*, without considering the often many possible alternative arrangements of bonds which leave these atoms connected.

It will be seen that the overall form of $S(N, m)$ can be characteristic of the molecular topology. In other words, the numbers and types of subgraphs of each size are dictated solely by the individual connectivities of the atoms. We will argue that the number of distinct subgraphs can be used as a guide to ''topological complexity,'' implying that the presence of high-connectivity atoms is an indication of a complicated topology. Consequently, our ability to compare topologies encompasses only graph-theoretical equivalence. Any deformation of a molecular graph that preserves adjacency relationships does not alter the topology.

To convert the purely topological information contained in the subgraphs into the chemical attributes of the parent molecule, the subgraphs must be compared with one another. There are many levels at which this may be carried out. Here the subgraphs are distinguished only by the atomic numbers of the atoms involved and the bond types of the bonds between them. Therefore, at this level, all carbon atoms are equivalent to one another, as are all oxygens, etc. It is only the types and numbers of bonds that they form which differentiates them. For example, a carbonyl carbon can only be distinguished from an aliphatic carbon in fragments which contain the carbonyl oxygen atom also. A substructure defined in this way will be referred to as a ''moiety'' to avoid confusion with other terms such as ''functional group'' or ''substructure.'' The term ''fragment'' is more general and can be used to mean either subgraph or moiety.

The distinct moieties are stored in different fields according to their size and it is, of course, only necessary to compare moieties of the same size. The method used to determine whether two moieties are identical is the subgraph isomorphism algorithm of Ullman.[19] Strictly, this is a problem in graph isomorphism because only graphs of the same size are being compared, but the Ullman algorithm is not restricted to subgraph isomorphisms and has been shown to be a particularly efficient method in general in chemistry.[20] When a match occurs, a moiety counter is incremented and the moiety in question is not stored again; if no match is made, the new moiety is appended to the file. It is usually true that the number of distinct moieties is much less than the number of distinct subgraphs for any molecule. The numbers of distinct moieties of size m for the molecule of N atoms is represented by the distribution function,

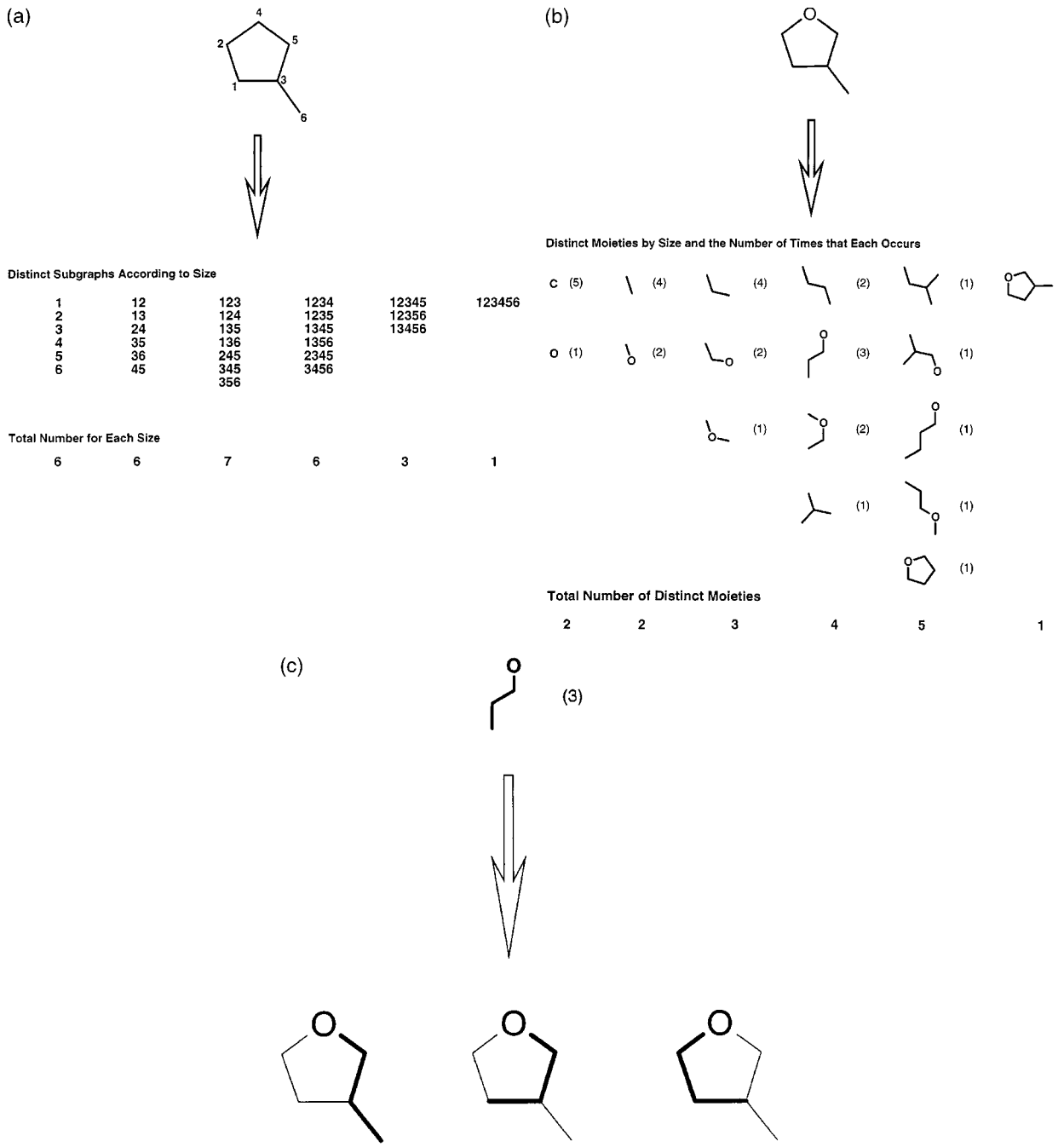

FIGURE 1. (a) Distinct subgraphs and their totals according to size for a methylcyclopentane skeleton. Subgraph numberings refer to atom labels on the skeleton. (b) Distinct moieties, their frequencies of occurrence, and the total number for each size for 3-methylfuran. (c) Occurrence of an example four-atom fragment from 3-methylfuran.

$M(N, m)$, where again m takes values between 1 and N. Consequently, $M(N, 1)$ is the number of distinct atom types in the molecule.

The distinct moieties for 3-methyltetrahydrofuran are shown in Figure 1b along with the number of times that each occurs. The numbers of distinct moieties for each respective size are shown at the bottom of the figure. In Figure 1c, a single moiety is shown along with its positions of occurrence in the original molecular skeleton.

The above ideas have been realized through a newly developed computer program, *RAPTOR*.[21] The input is the ''Mol.-file'' representation of molecular structure.[22] The output of the program

includes the numbers of distinct moieties and their respective frequencies for each molecule. The moieties are binned into separate files according to their size so that they may be straightforwardly analyzed with standard database software.[23] The calculations were performed on a desktop workstation.

Results

A systematic analysis of number of different topologies which are widespread in chemistry is presented and it is shown that, in each case, the distributions $S(N, m)$ and $M(N, m)$ are distinctive. The starting point is a study of the simplest forms, straight chains, branched systems, rings, and cages. This is followed by a consideration of rings with methyl substituents. An assessment of the effect of heteroatom substitution on each is then given. Then, consideration is given to more complicated topologies, exemplified by some fused ring system, before looking at amino acids as examples of naturally occurring molecules.

Before proceeding with the analysis, it is important to note that certain results are fundamental properties of connected graphs and therefore hold for all the molecules. The numbers of subgraphs of size 1 and 2 are obviously just the numbers of atoms and bonds, respectively, that is, $S(N, 1) = N$ and $S(N, 2) = E$. Furthermore, there can only be one subgraph whose size is the same as the parent molecule, that is, $S(N, N) = 1$. In addition, assuming that our molecular graphs may be rendered planar (i.e., may be drawn in such a way that no pairs of edges cross one another) the numbers of atoms and bonds are related by:

$$E = N - 1 + R \tag{1}$$

R is the ''cyclomatic number''[17] and defines the smallest number of rings from which all other rings may be constructed by linear combination.[24] Therefore, for a fixed number of atoms, introducing rings increases the number of bonds, a fact which has significant ramifications for the numbers of subgraphs and moieties present in cage molecules and fused ring systems.

Finally, the sums of all the atom connectivities, d, is equal to twice the number of bonds; that is:

$$\sum_{i=1}^{N} d_i = 2E = 2S(N, 2) \tag{2}$$

This relation is useful in the comparison of molecules with identical numbers of, respectively, atoms and bonds, because the average connectivities, $\bar{d}$, of their atoms are also the same. Thus, fair comparisons among sets of related molecules can be made, and the influence of individual connectivities on substructure distributions can also be demonstrated.

SIMPLE TOPOLOGIES

In this subsection, the distributions of subgraphs and moieties are presented for skeletons that contain only single atom types and in which all the bond types are equivalent.

Chains

For straight-chain molecules, the subgraphs are also all simple chains, ranging in size from single vertices up to the size, N, of the molecule itself. Their numbers are given by the simple linear relationship:

$$S(N, m) = N - m + 1 \tag{3}$$

These data are shown for straight-chain molecules up to length 12, in a figure that is available as supplementary material.

However, because all the atom types are the same and all the subgraphs are chains, there is no chemical distinction between any subgraph and any other of the same size. Therefore, there is only one distinct moiety for each size; that is, $M(N, m) = 1$ for all N and m, and linear molecules contain the least possible substructural diversity.

Branched Acyclic Systems

The effect of increasing the degree of branching may be assessed and compared with the simple relationships that were found for chains. The 75 isomers of decane were ranked according to the length of the longest chain of atoms present in each. The numbers of isomers with longest chains of length 10, 9, 8, 7, 6, and 5 are, respectively: 1, 4, 13, 26, 22, and 9. The averaged values of $S(N, m)$ and $M(N, m)$ at each m for these sets are shown in Figure 2a and b, respectively. Note that the ''degree of branching'' correlates with the shortness of the longest chain; that is, the isomers whose longest chain is of length 5 are the most highly branched.

It is clear from Figure 2a that the greater the degree of branching the more subgraphs of a given

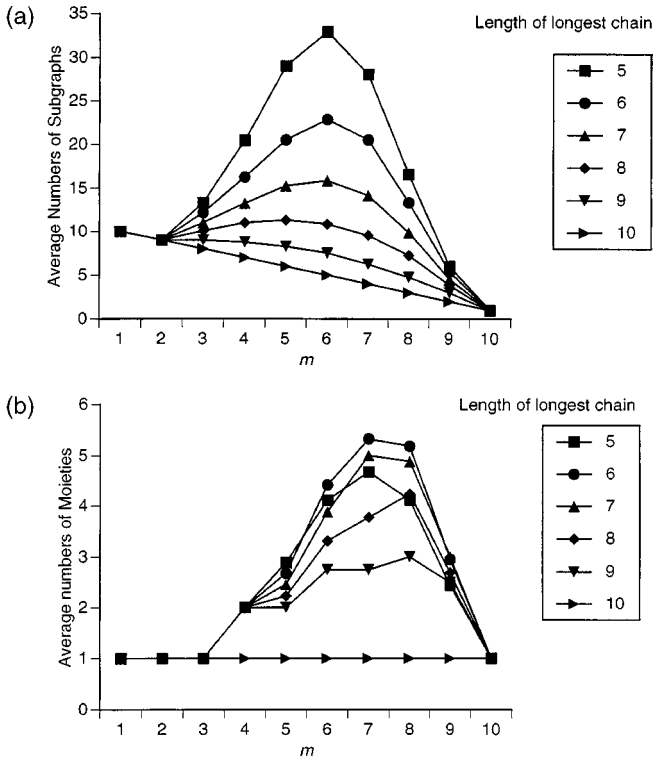

FIGURE 2. Fragment counts for decane isomers grouped by their longest chain lengths and averaged. (a) Subgraphs. (b) Moieties.

size that are generated on average. But branching introduces no new bonds relative to a straight chain with the same number of atoms. This means that the *average* atom connectivity for each molecule remains the same as that for a straight chain, that is, $2E/N = 2(N-1)/N$. A necessary consequence of shortening the length of the longest chain is that atoms with connectivity higher than the average are introduced. Therefore, the numbers of subgraphs of any size that can be generated is increased by the presence of high connectivity atoms.

Whereas for an unbranched straight chain the number of subgraphs falls off linearly with increasing subgraph size, for any branching at all the distributions are nonlinear and there may be a peak in each distribution situated at subgraph sizes larger than 2. [The initial decrease from $m = 1$ to $m = 2$ is because $S(N,2) = S(N,1) - 1$, i.e., $E = N - 1$, for acyclic molecules.] The presence of the peak in the distributions therefore represents a trade-off between two effects. The number of distinct topologies possible increases with increasing subgraph size; however, the number of ways in which any one of those topologies may be chosen from the skeleton generally decreases with increasing subgraph size. These factors also account for the observation that the more highly branched the molecule is, the larger is the size of subgraph that is most abundant. Alternatively, the more extended the molecule is, the more closely its distribution, $S(N,m)$, resembles that for the perfect straight chain in which the smallest subgraph is the most abundant.

The trends in the average moiety distributions (Fig. 2b) are broadly similar to those for the subgraphs in that the more highly branched the molecule the greater the number of distinct moieties of a given size that it contains. This is simply because the number of distinct topologies that will be found for each moiety size increases with the number of branched atoms.

Monocyclic Systems

For simple, unsubstituted, single-ring systems in which all bonds are equivalent, (e.g., cyclopropane, benzene, etc.) the numbers of subgraphs and moieties also follow very simple relations. For a ring of N atoms, there are exactly N subgraphs for each size m less than the ring-size itself, a result which is displayed in a figure for monocyclic systems with up to 12 atoms (refer to supplementary material). All such subgraphs are simple chains of length m, and for given m, they are identical with one another, so that, just as with simple chains, the number of distinct moieties is again exactly 1 for each size.

Comparison of the relationships for rings and chains is instructive. The form of $S(N,m)$ is qualitatively different for two classes of molecules. Although $S(N,m)$ for both series have a linear dependency on N for given m, for fixed N it is possible to derive many more subgraphs from the ring than from the chain. The presence of a ring allows chains of any length (up to the ring size) to be constructed starting from every atom in the ring whereas this is not possible for a chain. Note, though, that the difference in average atom connectivity, $\bar{d}$, between a ring (for which $\bar{d} = 2$) and a chain with the same number of atoms is small [for which $\bar{d} = 2(N-1)/N$]. In any case, unsubstituted rings offer no more in the way of substructural diversity than chains.

Cage Molecules

Cages illustrate the dramatic increase in subgraph and moiety counts with increasing average atom connectivity, $\bar{d}$, as shown in Table I. The

TABLE I.

Numbers of Distinct Subgraphs and Moieties for a Selection of Cage Molecules, Ordered by Average Atom Connectivity and Size. Subgraph Distributions Are Displayed in Figure 5 for Cages with Fewer than 12 Atoms.

		m																		
	$\bar{d}$	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
(a) S(N, m)																					
Adamantane	2.40	10	12	18	28	48	64	70	39	10	1										
Prismane	3.00	4	6	4	1																
Cubane	3.00	8	12	24	38	48	28	8	1												
Dodecahedrane	3.00	20	30	60	140	312	690	1480	3150	6500	12,588	21,540	30,470	32,700	24,620	12,672	4495	1120	190	20	1
C_{60}[a]	3.00	60	90	180	420	1032	2570	6480	16,380	41,560	106,212	272,640									
Octahedrane	4.00	6	12	20	15	6	1														
Cuboctahedrane	4.00	12	24	56	126	276	500	660	480	220	66	12	1								
$B_8H_8^{2-}$	4.25	8	18	40	65	56	28	8	1												
Icosahedrane	5.00	12	30	80	210	492	812	780	495	220	66	12	1								
(b) M(N, m)																					
Adamantane	2.40	1	1	1	2	2	4	5	4	2	1										
Prismane	3.00	1	1	1	1																
Cubane	3.00	1	1	1	3	2	3	1	1												
Dodecahedrane	3.00	1	1	1	2	3	5	8	15	24	46	75	139	211	211	118	53	14	5	1	1
$C_{60}H_{60}$[a]	3.00	1	1	1	2	3	6	10	21	41	94	196									
C_{60}[a]	3.00	1	2	2	5	8	19	34	84	181	436	990									
Octahedrane	4.00	1	1	2	2	1	1														
Cuboctahedrane	4.00	1	1	2	3	5	9	13	14	9	4	1	1								
$B_8H_8^{2-}$	4.25	1	1	2	4	7	7	2	1												
Icosahedrane	5.00	1	1	2	3	6	13	11	10	5	3	1	1								

[a]Only subgraphs and moieties up to size 11 could be computed for C_{60} skeletons.

distributions, $S(N, m)$, for a selection of these molecules with $N \leq 12$ are shown in a figure in the supplementary material. For all the molecules, the distributions $S(N, m)$ each have a single peak and, for given N, the numbers of subgraphs are much greater than for either rings or chains.

Consider a progression of molecules of different sizes whose carbon atoms all have a connectivity of 3: C_4H_4 (''prismane''); C_8H_8 (''cubane''); $C_{20}H_{20}$ (''dodecahedrane''); and C_{60} (''Buckminster Fullerene''). (Due to computational limitations, it has only been possible to compute the subgraph up to size 11 for C_{60}.) The numbers of subgraphs for these topologies escalate with m much more rapidly than was possible for either linear, branched, or cyclic systems. For C_{60}, they follow an approximately geometric progression, $S(60, m) \approx 2.55\ S(60, m-1)$, for those shown $(3 < m \leq 11)$.

Considering now molecules which are of the same size but which have different $\bar{d}$, more subgraphs of a given size are generated for those molecules with higher $\bar{d}$; for instance, compare cubane with $B_8H_8^{2-}$ (both have 8 atoms) and ''cuboctahedrane'' with icosahedrane, (both have 12 atoms).

The distributions in numbers of moieties for the cages follow broadly similar trends to those for subgraphs except that the distributions are much flatter overall and may have two maxima, for example, for adamantane and cubane. The actual degeneracy of the moieties of these skeletons is very high, due to their symmetry.[25] It is clear that cage molecules give rise to significant substructural diversity and therefore would form useful components of a chemical library.

Summary of Simple Topologies

The main conclusion to be drawn from this analysis of simple skeletons applies to the comparison of molecules with equal average connectivities, $\bar{d}$. The presence of a small number of high-connectivity atoms can increase the numbers of subgraphs significantly relative to a molecule whose atoms mostly have average connectivities. But, while an increase in $\bar{d}$, for a fixed number of atoms, can also lead to an increase in the number of subgraphs it will not always exceed the number that can be obtained by introduction of branching points. The branched decanes (for which $\bar{d} = 1.8$) demonstrated that it was possible to obtain more subgraphs for some isomers than from cyclodecane ($\bar{d} = 2.0$), although fewer than adamantane ($\bar{d} = 2.4$).

INFLUENCE OF METHYL SUBSTITUENTS ON SIMPLE RINGS

Simple monocyclic rings with single methyl substituents and pairs of them in 1,2-, 1,3-, 1,4-, and 1,5- relationships are considered. These systems contain extra atoms and bonds with respect to the simple rings, so direct comparison between molecules whose rings are the same size is not reasonable. Subgraph and moiety distributions for the 1,4-substituted ring molecules with up to 12 atoms are shown in Figure 3. The most obvious feature is the sharp peak in the subgraph distribution, $S(N, m)$, which is much more skewed toward large fragment size than the peak in $S(N, m)$ for the cages and the branched decanes.

Each substituent has the effect of increasing the connectivity of each ring atom to which it is bonded while introducing a ''terminal'' atom so the num-

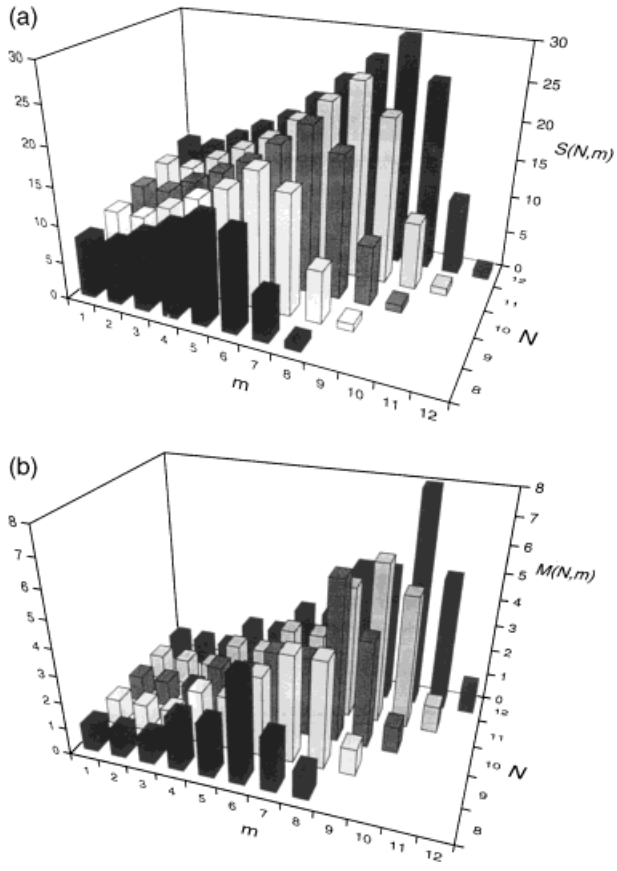

FIGURE 3. Fragment counts for each size, m, for 1,4-dimethyl rings with up to 12 atoms. (a) Subgraphs. (b) Moieties.

ber of bonds is still the same as the number of atoms and the average connectivity remains the same as for an unsubstituted ring (i.e., $\bar{d} = 2.0$). As has already been understood from the branched decanes, the presence of even small numbers of atoms with higher connectivities than the average leads to significant increases in both moiety and subgraph counts. Consequently, there are more subgraphs of size m compared with unsubstituted rings of the same size, N. This is easily explained for a singly substituted ring with N atoms. The ring is of size $N - 1$ and there will be $N - 1$ subgraphs of size m which comprise chains of ring atoms, but there will also be $m - 1$ subgraphs which contain $m - 1$ connected ring atoms and the substituent atom itself, that is, the total is $N + m - 2$ subgraphs, to be compared with N for an unsubstituted ring of N atoms. Similar criteria apply to the doubly substituted rings.

Algebraic relationships for the subgraph distributions, $S(N, m)$, are summarised in Table II for single substitutions and different double substitution patterns. Note that N refers to the number of atoms in total and not just the number of atoms in the ring. The relationships all show a linear dependence on N and m and, for intermediate fragment sizes, the progressions parallel one another. It can be seen that, for the doubly substituted rings, although the subgraph distributions are very similar to one another, slightly fewer subgraphs are found for given size, m, as the substituents move further apart from one another. This is a subtle effect, due to the fact that, as the substituents move further apart, the number of subgraphs which contain both atoms with three-fold connectivity decreases. Increased branching leads to an increase in the number of subgraphs of a given size. When the two substituted atoms are closest to one another, the number of different branching possibilities for subgraphs which include them is the highest.

The moiety counts, exemplified in Table III for substituted 10-membered rings, exhibit the same basic trends and vary little with the substitution pattern. The total number of distinct moieties for each of the doubly substituted rings also decreases as the substituents move further apart. $M(N, m)$ for the substituted rings consistently peaks at size $N - 1$ for the single substituted rings and $N - 2$ for the doubly substituted rings. This positioning of the peak can be explained by using singly substituted rings as an example. For any size ($N - 1 > m \geq 2$) there will be only one distinct moiety which comprises a chain of ring atoms or a chain which uses the substituent as one of its endpoints. But additional, distinct moieties may be composed of chains of length $m - 1$, with the substituent atom at points along their length other than the two terminal positions. Even taking account of symmetry, the number of distinct moieties that are branched is greatest at size $N - 2$. For size $m = N - 1$, there is a single moiety which is the complete ring itself, a chain of length $N - 1$ using the substituent atom as one of its endpoints and chains of length $N - 2$ with the single substituent at points along their length. Thus, the greatest number of distinct moieties occurs at $m = N - 1$.

Therefore, introduction of substituents into simple rings can increase the moiety count significantly and in such a way that variety is greatest for the largest fragment sizes.

INFLUENCE OF SUBSTITUTION OF SINGLE TYPES OF HETEROATOM ON SIMPLE TOPOLOGIES

In this section, the effect of introducing heteroatoms into the basic skeletons already described is considered. While this action does not

TABLE II.
Algebraic Relationships for Numbers of Distinct Subgraphs for Simple Monocyclic Systems with Single and Double Substituents, for $N \leq 22$.

Substitution	$S(N, m)$ 1	2	3	4	5	m	$N - 3$	$N - 2$	$N - 1$	N
None						N				1
1	N	N				$N + m - 2$			$N - 1$	1
1,2	N	N	$N + 2$			$N + 3m - 7$		$3N - 10$	$N - 2$	1
1,3	N	N	$N + 2$			$N + 3m - 8$		$3N - 11$	$N - 2$	1
1,4	N	N	$N + 2$	$N + 4$		$N + 3m - 9$		$3N - 11$	$N - 2$	1
1,5	N	N	$N + 2$	$N + 4$	$N + 4$	$N + 3m - 10$	$N + 3m - 9$	$3N - 11$	$N - 2$	1

TABLE III.
Numbers of Distinct Moieties, $M(N, m)$, for Singly and Doubly Substituted 10-Carbon Rings.

	m												
Substitution	1	2	3	4	5	6	7	8	9	10	11	12	Total
None	1	1	1	1	1	1	1	1	1	1			10
1	1	1	1	2	2	3	3	4	4	6	1		29
1,2	1	1	1	2	2	4	4	6	6	9	5	1	42
1,3	1	1	1	2	2	3	4	5	6	8	6	1	39
1,4	1	1	1	2	2	3	3	5	5	8	5	1	37
1,5	1	1	1	2	2	3	3	4	5	7	6	1	36

alter the subgraph counts, there are chemical consequences; that is, the number of distinct moieties varies significantly. Therefore, the topic of discussion in this subsection is largely restricted to the moiety distributions, $M(N, m)$.

Straight-Chain Alcohols, Peroxides, and Ethers

Single atoms of a carbon chain may be replaced with an oxygen atom. Substitution at the terminal atom gives rise to a sequence of primary alcohols. Replacement of carbon atoms in the 2-, 3-, 4-, and 5- positions generates series of methyl, ethyl, *n*-propyl, and *n*-butyl ethers, respectively. Of course, none of these distributions depends on the nature of the heteroatom itself; therefore, analogous series of thiols and thio-ethers, for example, give rise to identical distributions of moieties.

The progression of moiety counts for altering the position of the single oxygen atom in a series of 12-atom chains is such that more distinct moieties of any length are obtained, the closer the heteroatom is to the center of the molecule. In addition, as the heteroatom moves closer to the center of the molecule, the moiety distributions become peaked in their centers. A figure representing this data is available as supplementary material. Figure 4 shows the way the distribution of moieties changes with increasing chain length for chains substituted at the fourth position.

It is also possible to make double substitutions of the basic chains. A representative set may be obtained by keeping one substituent fixed at a terminal position and moving the other to positions, 2, 3, 4, and 5. (In this way, 1,2- and 1,3-substitutions of oxygen atoms lead to peroxides and hemiacetals respectively.) In a similar approach to that shown for the singly substituted chains, the effect of altering the substitution pattern on a fixed chain length and the distributions of moiety counts

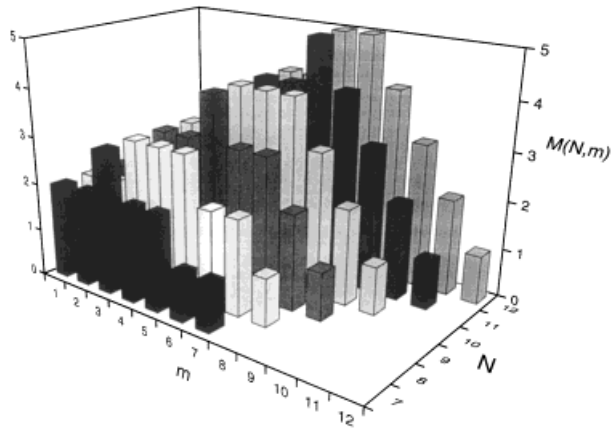

FIGURE 4. Numbers of distinct moieties, $M(N, m)$, for each size m, for straight-chain propyl-ethers up to 12 atoms.

with increasing chain length for the 1,4-substituted chains have been considered. These data are available as supplementary material.

There is very little difference in the moiety counts between the singly substituted and doubly substituted chains. The only difference of note is that there are more distinct moieties for small m in the doubly substituted than in the corresponding singly substituted examples. The peak moiety count reaches the same upper limit, for given N, whether the terminal atom is substituted or not. This demonstrates that the number of distinct moieties that may be obtained is severely limited by the topology.

Heptamines

The nine isomers of heptane (Fig. 5) serve as templates here; subgraph and moiety counts for each are shown in Table IV. To investigate the effect of single heteroatom substitution on

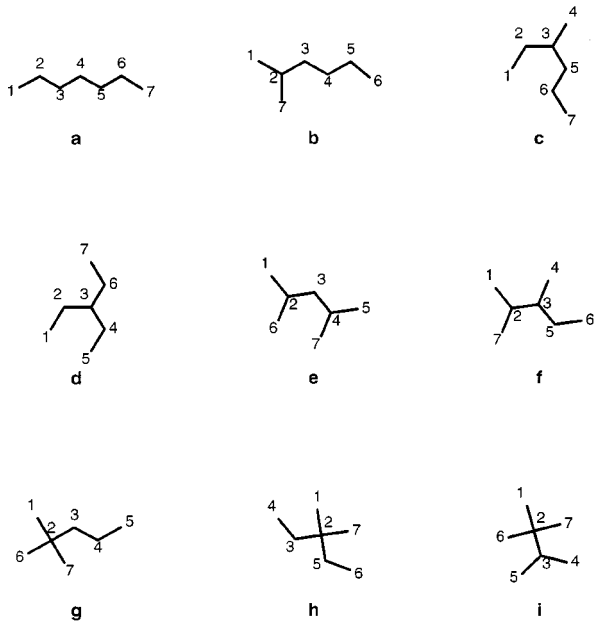

FIGURE 5. Heptane skeletons. Atom labels denote symmetry-unique positions substituted by nitrogen in the study of amines.

branched acyclic systems, each symmetry-unique carbon of the eight branched isomers of heptane is replaced with a nitrogen atom, which leads to examples of primary, secondary, and tertiary amines and quaternary ammonium derivatives (Table V). (Substitutions of *n*-heptane, **a**, are omitted because the results would not be distinguishable from those for the alcohols and ethers in the previous subsection). The molecules are grouped in the table according to their skeletal form (in descending order of longest chain) in the same order of the heptane isomers in Table IV; their numbering corresponds with the number of the carbon atom substituted in each case (Fig. 5).

The numbers of distinct moieties for the amines are greater than that for the respective heptanes and there is a slight increase in the numbers of fragments of size 4, 5, and 6 with increased branching. The fact that introduction of a heteroatom increases the number of moieties is in line with expectation, but the distributions are rather flat overall and the underlying topology remains the dominant factor.

TABLE IV.
Subgraph and Moiety Counts for Acyclic Heptanes (see Fig. 11).

		Size (m)						
		1	2	3	4	5	6	7
$S(7, m)$								
a	*n*-heptane	7	6	5	4	3	2	1
b	2-methylhexane	7	6	6	5	4	3	1
c	3-methylhexane	7	6	6	6	5	3	1
d	3-ethylpentane	7	6	6	7	6	3	1
e	2,2-dimethylpentane	7	6	8	8	7	4	1
f	2,3-dimethylpentane	7	6	7	8	7	4	1
g	2,4-dimethylpentane	7	6	7	6	6	4	1
h	3,3-dimethylpentane	7	6	8	10	8	4	1
i	2,2,3-trimethylbutane	7	6	9	11	10	5	1
$M(7, m)$								
a	*n*-heptane	1	1	1	1	1	1	1
b	2-methylhexane	1	1	1	2	2	2	1
c	3-methylhexane	1	1	1	2	2	3	1
d	3-ethylpentane	1	1	1	2	2	1	1
e	2,2-dimethylpentane	1	1	1	2	3	2	1
f	2,3-dimethylpentane	1	1	1	2	2	3	1
g	2,4-dimethylpentane	1	1	1	2	2	1	1
h	3,3-dimethylpentane	1	1	1	2	3	2	1
i	2,2,3-trimethylbutane	1	1	1	2	2	2	1

TABLE V.
Moiety Counts, $M(7, m)$, for Branched Heptamine Isomers, Labeled b – i According to Skeletons in Figure 11, Where Number Denotes Position of N Atom in Each Respective Skeleton.

Structure / N position		1	2	3	4	5	6	7
b	**1, 7**	2	2	2	3	3	3	1
b	**2**	2	2	3	4	3	2	1
b	**3**	2	2	3	3	3	2	1
b	**4**	2	2	3	3	3	2	1
b	**5**	2	2	3	4	3	2	1
b	**6**	2	2	2	3	3	2	1
c	**1**	2	2	2	3	4	3	1
c	**2**	2	2	3	4	5	3	1
c	**3**	2	2	3	3	3	3	1
c	**4**	2	2	2	3	3	3	1
c	**5**	2	2	3	4	4	3	1
c	**6**	2	2	3	4	4	3	1
c	**7**	2	2	2	3	3	3	1
d	**1, 5, 7**	2	2	2	3	4	2	1
d	**2, 4, 6**	2	2	3	4	4	2	1
d	**3**	2	2	2	2	2	1	1
e	**1, 5, 6, 7**	2	2	2	4	4	3	1
e	**2, 4**	2	2	3	4	3	2	1
e	**3**	2	2	3	2	2	1	1
f	**1, 7**	2	2	2	4	5	4	1
f	**2**	2	2	3	5	4	3	1
f	**3**	2	2	3	3	3	3	1
f	**4**	2	2	2	4	4	3	1
f	**5**	2	2	3	5	5	3	1
f	**6**	2	2	2	3	3	3	1
g	**1, 6, 7**	2	2	2	4	5	3	1
g	**2**	2	2	3	3	3	2	1
g	**3**	2	2	3	3	3	2	1
g	**4**	2	2	3	3	3	2	1
g	**5**	2	2	2	3	3	2	1
h	**1, 7**	2	2	2	4	4	3	1
h	**2**	2	2	2	2	3	2	1
h	**3, 5**	2	2	3	5	5	3	1
h	**4, 6**	2	2	2	3	4	3	1
i	**1, 6, 7**	2	2	2	4	4	3	1
i	**4, 5**	2	2	2	4	4	3	1
i	**2**	2	2	3	3	3	2	1
i	**3**	2	2	3	4	3	2	1

Cyclic Ethers

In the same way, the effect of heteroatom substitution on the simple monocyclic rings can be investigated. For this series there is only one distinct way to make a single heteroatom substitution; therefore, double substitutions in 1,2-, 1,3-, 1,4-, and 1,5- relationships are also considered.

The moiety counts for the 1,4-substituted molecules and those of ring size 12 with different substitution patterns are displayed graphically in Figures 6 and 7, respectively. Overall, a greater number of distinct moieties is possible for a given substitution pattern in the ring than in the chain. As described previously, this is due to termination

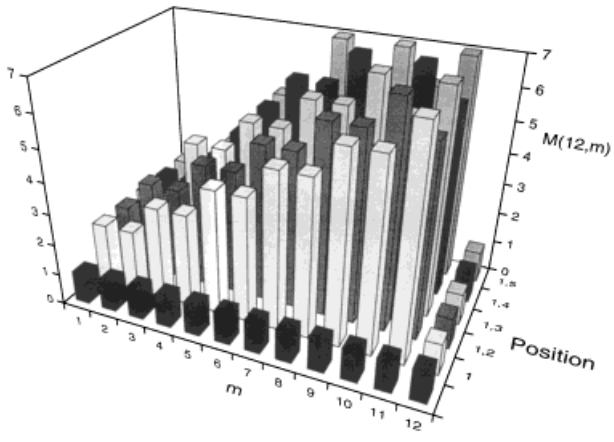

FIGURE 6. Numbers of distinct moieties of each size, *m*, for oxygen-substituted monocyclic rings of size 12 ordered by position of substitution: none; 1: 1,2 (cyclic peroxide); 1,3; 1,4; and 1,5.

effects which truncate the fragment distribution for chain molecules. A given sized subgraph (which is a chain) may be obtained by starting at every atom in the ring, so more are possible (for sizes $m > 1$) than for straight-chain molecules. Hence, chains of the same length that come from the rings may contain the substituted heteroatoms at a greater number of positions than those from the straight-chain molecules, and, therefore, give rise to many more distinct moieties. A consequence is that for a given substitution pattern, the numbers of distinct moieties of intermediate sizes which are obtained from the chains will level off and become independent of chain size, whereas for the rings, the number increases progressively as the ring size increases.

Furthermore, for a given N there are approximately the same number of distinct moieties from the singly substituted cyclic ethers (Fig. 7) as with the single methyl-substituted rings (Table III), but the cyclic di-ethers have slightly fewer distinct moieties than the comparable double methyl-substituted rings. This demonstrates, ultimately, the crucial role that branching plays in creating substructural diversity, because significant increases in moiety counts may be achieved with branching alone, without any recourse to heteroatom substitution.

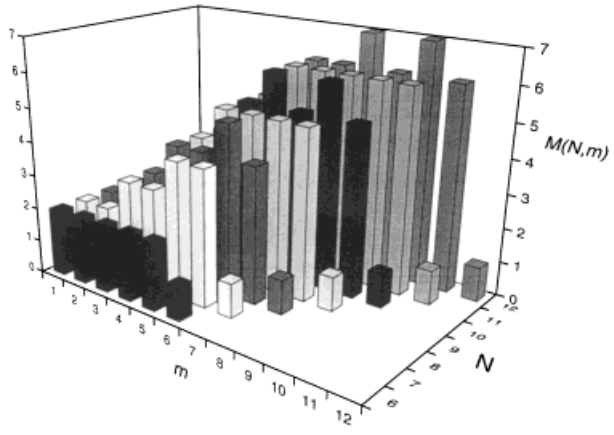

FIGURE 7. Numbers of distinct moieties of each size, *m*, for monocyclic rings with up to 12 atoms with oxygen substituents in first and fourth positions.

Substituted Cages

The last category of molecules in this subsection is the substituted icosahedral skeleton. The plain skeleton represents $B_{12}H_{12}$; a ''singly substituted'' carborane is created by replacing one BH moiety by a carbon atom. Similarly, three more distinct carboranes may be created by the three symmetrically distinct ways in which two carbons may be positioned in the skeleton, *viz*. in 1,2-, 1,3-, and 1,7-relationships, according to the usual numbering scheme.

Table VI summarizes the moiety distributions. For single substitution alone, there is a pronounced effect, and the number of distinct moieties for each subgraph size is four to five times as many as for the unsubstituted skeleton. For the doubly substituted isomers, symmetry has an influence. The 1,2- and 1,3-isomers, both with C_{2v}

TABLE VI.
Counts of Distinct Moieties in Singly and Doubly Substituted Icosahedral Cages.

m	1	2	3	4	5	6	7	8	9	10	11	12
Icosahedron	1	1	2	3	6	13	11	10	5	3	1	1
1-Carborane	2	2	5	10	24	57	65	52	28	12	4	1
1,2-Dicarborane	2	3	7	17	42	105	133	112	59	23	5	1
1,3-Dicarborane	2	2	6	14	39	105	133	112	59	23	5	1
1,7-Dicarborane	2	2	5	11	24	49	48	37	16	8	2	1

symmetry,[25] have very similar profiles, differing only in the numbers of distinct moieties of the smaller sizes. The 1,7-isomers has D_{5h} symmetry and, consequently, the number of distinct moieties is less than for the other doubly substituted isomers. On the whole, it also has fewer distinct substructures than the singly substituted isomer which has C_{5v} symmetry.

The influence of bond type may also be considered, illustrated by two C_{60} skeletons. Moiety counts for a C_{60} skeleton in which all bond types are equivalent, for example, the C—C single bonds in a $C_{60}H_{60}$ molecule may be compared with a ''Kekulé'' form in which alternating double and single bonds cover the framework. The latter is consistent with the accepted bonding patterns in C_{60} itself, in which C—C bonds shared by two hexagons differ in length from those shared by a hexagon and a pentagon. The difference in moiety counts (Table I), despite the high symmetry[25] of both forms, is quite significant—there being approximately five- to six-fold increase in the numbers of distinct moieties in the unsaturated form, for $m > 8$.

Summary of Heteroatom Substitution Effects

Heteroatom substitution in simple topologies leads to marked substructural diversity with respect to the unsubstituted skeletons. Rings now give rise to a greater number of distinct moieties than straight chains with the same number of atoms. Geometric symmetry of the skeleton,[25] where present, plays a significant role in determining the number of distinct moieties, but the limiting factor in the generation of distinct moieties in each case is the underlying topology. Therefore, for single heteroatom substitution, the molecules which give rise to the greatest number of subgraphs also give rise to the greatest number of distinct moieties.

FUSED-RING SYSTEMS

The analysis is now extended to a treatment of fused polyhexes. These topologies are more complicated than the simple systems described above, and are widespread in organic chemistry, often forming the basic core structure of therapeutic agents.

Fused-ring systems illustrate the sensitivity of the distinct moiety counts to geometric symmetry of the parent molecule,[25] even when all constituent atoms are identical. Two series of edge-fused polyhexes are considered. It is assumed that all of the bond types are equivalent. The fused polyhexes are conveniently categorized by their numbers of atoms and bonds, that is, the pair, $\{N, E\}$. For these systems, all molecules in the same category have exactly the same number of atoms of a given connectivity. The $\{18, 21\}$ set of molecules, which each have exactly four six-membered rings, is shown in Fig. 8a). Each molecule has 6 atoms with connectivity 3 and 12 atoms with connectivity 2. Note that in this set no atom is shared by more than two six-membered rings. The $\{21, 25\}$ set, comprising molecules with exactly five six-membered rings is also considered (Fig. 8). In this series, all molecules have 8 atoms with connectivity 3 and 13 atoms with connectivity 2 and, for any molecule, exactly one atom is shared between three six-membered rings.

It is clear from Table VII that $S(N, m)$ for members within each of the two series are very similar. In fact, the greatest differences appear only for subgraphs in the middle of the size range. In the $\{18, 21\}$ series, fragments of size 5 have to be counted before the differences between the five molecules can be resolved; in the $\{21, 25\}$ series, size 7 has to be counted. Note also, that for the larger fragments, closer to the size of the parent molecule, there are also very fine distinctions. In fact, individual members of the $\{18, 21\}$ and the $\{21, 25\}$ series of molecules may only be fully distinguished for subgraphs smaller than size 14 and 17, respectively.

The form of the distributions, overall, is quite similar to that for cages, in that there is a single prominent peak. Furthermore, the number of subgraphs (and moieties) of any size greatly exceeds that for linear or monocyclic molecules with the same number of atoms. The presence of small numbers of atoms with $d = 3$, and the fact that any pair of them is separated by at most four atoms are responsible for this. As was seen for the dimethyl-substituted rings, the more atoms to be found in close proximity to highly branched centers, the greater the level of branching within any subgraph and the greater the number of subgraphs possible.

The form of $M(N, m)$ is similar to $S(N, m)$ in each case but the influence of symmetry on comparisons between the forms[25] is seen somewhat starkly. Two measures of graph symmetry are considered. First, there is the number of elements in the automorphism group of each graph. This quantity denotes the number of distinct permutations of labeled graph vertices which preserve all adja-

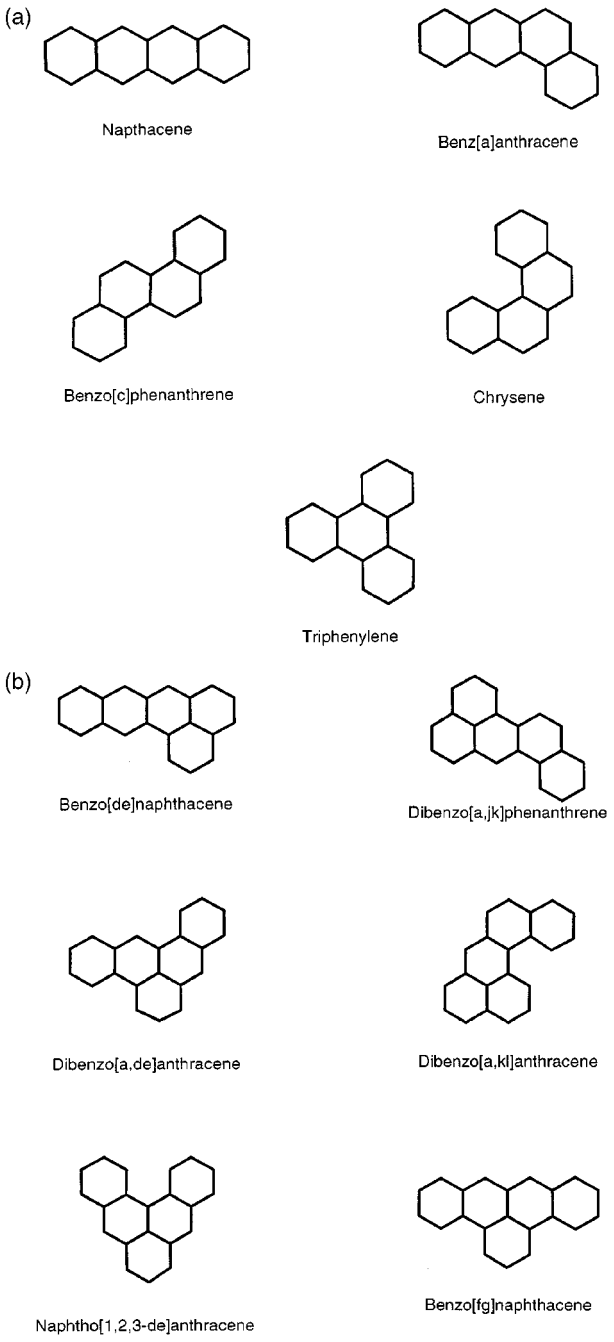

FIGURE 8. Structures of fused ring molecules containing only six-membered rings. (a) The set with 18 atoms and 21 bonds. (b) The set with 21 atoms and 25 bonds.

cency relationships. Second, we may consider classifying the nodes of the graph into classes according to their environments. It can be argued that the smaller is the number of classes the higher is the symmetry of the graph. If the two series of molecules are placed in ascending order of symmetry of their frameworks (according to these measures), then one obtains the ordering of Table VIII. For the {18, 21} set, the fact that the peak values for naphthacene (90) and triphenylene (81), the two most symmetric molecules, are less than half the peak values for the others is conspicuous. Benz[*a*]anthracene, the least symmetric, has the largest (or equal largest) number of distinct moieties for any value of m.

Among the {21, 25} set, the four molecules of C_s symmetry have very similar values across all fragment sizes, whereas the two molecules with C_{2v} symmetry have much lower moiety counts, overall. Note that, from the point of view of counting the moieties present, one has to enumerate all the way to $m = 10$ to differentiate all members of the {18, 21} set and to $m = 12$ for the {21, 25} set. These examples demonstrate convincingly the need to count fairly sizeable fragments to uncover topological distinctions between similar molecules.

AMINO ACIDS

The amino acids are provided as examples of important naturally occurring compounds whose substructure distribution is of general interest and in which the presence of several different heteroatoms heavily influences the number of distinct moieties.

Fragment distributions of the 20 naturally occurring amino acids were obtained. The molecules range in size from 5 heavy atoms (glycine) to 15 (tryptophane). In general, the average connectivity of the atoms is low although the skeletons have a ''branched'' topology. Table IX gives the distribution, $M(N, m)$, for the 20 molecules, in ascending order of the number of heavy atoms. $S(N, m)$ is supplied as supplementary material. A number of the amino acids are isoskeletal and so have identical entries for $S(N, m)$. These include: cysteine and serine; threonine and valine; asparagine, aspartine, and leucine; and glutamic acid and glutamine.

$S(N, m)$ has two different overall forms, depending on whether or not the molecule has a ring. The distributions of the acyclic molecules (alanine, cysteine, threonine, asparagine, isoleucine, methionine, glutamic acid, lysine, and arginine—and their respective isoskeletal partners) do not follow a regular increase in m to a single peak. Such molecules have more atoms ($m = 1$) than bonds ($m = 2$) so that $S(N, m)$ initially falls with increasing m, but there is a subsequent peak in the distribution at larger m. Similar forms are observed for the branched decane (Fig. 2), and ap-

TABLE VII.

Numbers of Distinct Subgraphs and Moieties for Two Series of Fused-Ring Compounds with Constant Numbers of Atoms and Bonds. Structures Are Shown in Figure 14.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
(a) $S(N, m)$																					
Naphthacene	18	21	30	48	82	127	188	278	392	533	692	796	780	619	346	114	18	1			
Benz[*a*]anthracene	18	21	30	49	86	138	214	325	468	638	806	896	840	635	346	114	18	1			
Benzo[*c*]phenanthrene	18	21	30	50	91	154	255	406	599	812	985	1036	916	651	346	114	18	1			
Chrysene	18	21	30	50	90	148	236	362	524	718	896	980	900	651	346	114	18	1			
Triphenylene	18	21	30	51	96	169	294	474	686	882	990	987	852	618	346	114	18	1			
Benzo[*de*]naphthacene	21	25	37	63	117	203	343	566	890	1356	1991	2731	3424	3808	3602	2806	1708	712	172	21	1
Dibenzo[*a*, *jk*]phenanthrene	21	25	37	64	121	214	370	620	993	1538	2268	3099	3833	4163	3846	2919	1720	712	172	21	1
Dibenz[*a*, *de*]anthracene	21	25	37	64	122	220	393	682	1114	1709	2436	3178	3768	3992	3658	2809	1694	705	172	21	1
Dibenz[*a*, *kl*]anthracene	21	25	37	64	122	219	386	662	1080	1689	2489	3360	4078	4338	3926	2935	1720	712	172	21	1
Naphth[1,2,3-*de*]anthracene	21	25	37	64	123	226	414	739	1229	1891	2666	3395	3918	4058	3665	2797	1694	705	172	21	1
Benzo[*fg*]naphthacene	21	25	37	64	122	219	387	661	1066	1633	2355	3141	3802	4079	3739	2837	1694	705	172	21	1
(b) $M(N, m)$																					
Naphthacene	1	1	1	2	2	5	7	13	18	31	43	68	86	90	66	31	5	1			
Benz[*a*]anthracene	1	1	1	2	2	5	7	15	25	55	94	168	231	269	220	103	18	1			
Benzo[*c*]phenanthrene	1	1	1	2	2	5	7	15	25	54	93	164	217	213	141	57	10	1			
Chrysene	1	1	1	2	2	5	7	15	25	52	83	139	178	190	134	57	9	1			
Triphenylene	1	1	1	2	2	5	7	14	23	44	64	94	94	81	47	20	3	1			
Benzo[*de*]naphthacene	1	1	1	2	2	5	7	15	26	59	109	230	421	734	1052	1216	1018	543	153	21	1
Dibenzo[*a*, *jk*]phenanthrene	1	1	1	2	2	5	7	15	26	59	110	232	417	714	1023	1198	1025	563	161	21	1
Dibenzo[*a*, *de*]anthracene	1	1	1	2	2	5	7	15	26	59	112	244	456	787	1090	1211	957	513	148	21	1
Dibenzo[*a*, *kl*]anthracene	1	1	1	2	2	5	7	15	26	59	111	243	458	815	1174	1326	1062	558	154	21	1
Naphtho[1,2,3-*de*]anthracene	1	1	1	2	2	5	7	15	26	58	109	227	399	603	705	671	501	269	81	12	1
Benzo[*fg*]naphthacene	1	1	1	2	2	5	7	15	26	59	110	229	391	615	779	804	608	309	87	12	1

TABLE VIII.
Symmetry of Fused-Ring Systems (See Fig. 8).

	h_{AUT}[a]	n_{class}[b]
{18, 21} Set		
Benz[*a*]anthracene	1	18
Benzo[*c*]phenanthrene	2	10
Chrysene	2	9
Naphthacene	4	5
Triphenylene	6	3
{21, 25} Set		
Benzo[*de*]naphthacene	1	21
Dibenzo[*a*, *jk*]phenanthrene	1	21
Dibenzo[*a*, *de*]anthracene	1	21
Dibenzo[*a*, *kl*]anthracene	1	21
Naphtho[1,23-*de*]anthracene	2	12
Naphtho[1,2,3-*de*]anthracene	2	12

[a] Order of the automorphism group.
[b] Number of classes of equivalent nodes.

pear to be a common feature of branched acyclic molecules. The amino acids which contain a ring (i.e., proline, histidine, phenylalanine, tyrosine, and tryptophane) do not have such a form for $S(N, m)$, but their subgraph distributions show a simple rise and fall with increasing m.

In Table IX, the influence of the presence of heteroatoms on the moiety counts is clearly observed. The presence of more than one different type of heteroatom means that there are many more distinct moieties than are possible in isoskeletal hydrocarbons. Moreover, the larger the fragment size, the closer is the number of distinct moieties to the number of subgraphs themselves. In the limiting case, in which every atom in a skeleton is distinct, $S(N, m)$ and $M(N, m)$ would be equal for all m.

Also conspicuous is the small total number of distinct moieties, that is, chemical substructures, which are present among the set of amino acids. The moieties of all sizes number a few hundred at most, to be compared with over 1000 which could be obtained (e.g., cages or fused polyhexes of comparable sizes). A ''binning'' exercise was carried out in which the 20 amino acids were pooled and the total numbers of distinct moieties of each size from the whole set were counted. They are histogrammed in Figure 9. Note that the numbers of substructures which can be obtained from the largest molecule, tryptophane, correspond to a significant fraction of the total number of distinct moieties among all the amino acids. For smaller sizes, approximately half of the moieties are present in tryptophane (although many of these can also be found in the other amino acids); and, for moieties larger than size 8, the fraction that can be

TABLE IX.
Numbers of Distinct Moieties, $M(N, m)$, for the 20 Naturally Occurring Amino Acids, Ordered by Size.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Glycine	3	4	4	3	1										
Alanine	3	4	5	6	4	1									
Cysteine	4	5	6	8	7	4	1								
Serine	3	4	5	6	6	4	1								
Proline	3	4	6	9	13	12	6	1							
Threonine	3	4	5	9	11	10	5	1							
Valine	3	4	5	9	9	7	4	1							
Asparagine	3	4	6	9	11	12	10	5	1						
Aspartic acid	3	4	5	8	9	10	8	5	1						
Isoleucine	3	4	5	9	11	12	10	5	1						
Leucine	3	4	5	9	10	9	7	4	1						
Methionine	4	5	7	10	10	9	7	4	1						
Glutamine	3	4	6	9	10	11	12	10	5	1					
Glutamic acid	3	4	5	8	9	9	10	8	5	1					
Lysine	3	4	5	8	9	9	9	7	4	1					
Histidine	3	6	10	16	19	21	25	25	18	7	1				
Arginine	3	5	8	12	13	13	12	12	12	10	5	1			
Phenylalanine	3	5	7	12	14	17	18	20	19	14	6	1			
Tyrosine	3	5	8	14	16	21	22	25	27	25	16	6	1		
Tryptophane	3	6	9	17	27	42	63	80	94	104	101	77	40	11	1

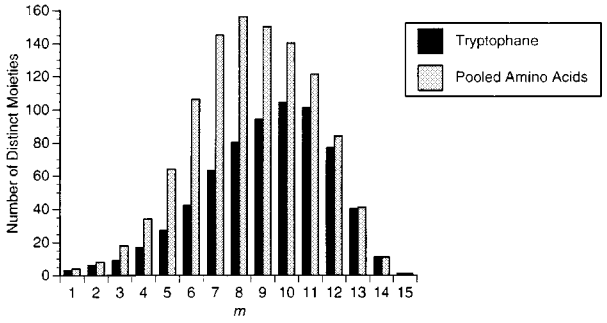

FIGURE 9. Histogram of numbers of distinct moieties for the amino acids. The gray bar represents the number of distinct moieties for all 20 molecules. The black bar represents the number of distinct moieties from tryptophane alone

found in tryptophane increases rapidly until size 15, which is attributable exclusively to it. The significantly larger number of moieties that tryptophane contains relative to any other amino acid may account for its overrepresentation at the binding sites of antibodies and other proteins.[27] It can be argued that the large number of substructures present in it afford a larger assortment of possible interactions with ligands than do the smaller amino acids.

COMPARISON OF ALL TOPOLOGIES

It is useful to conclude with a quantitative comparison of $S(N, m)$ and $M(N, m)$ for a range of topologies for fixed molecule size. Subgraph and moiety distributions for a selection of 12 atom systems are displayed in Figures 10 and 11 respectively. Not only do these figures reveal the utility of substructure enumeration for quantifying molecular topology, but they show that moiety counts are very sensitive to heteroatom substitution. Individual amino acid molecules are not very distinctive at the subgraph-counting level and rank among simple branched and cyclic structures. On the other hand, they stand out among the moiety counts, because they have significant heteroatom diversity. This is reinforced by the tendency for their skeletons to be branched. It remains clear that cagelike and fused-ring systems, especially those with low symmetry,[25] yield the greatest numbers of subgraphs and moieties for a given size and composition.

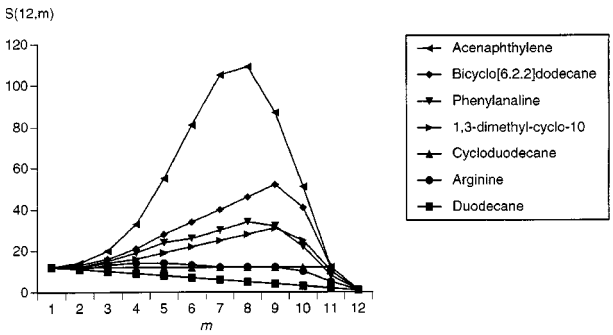

FIGURE 10. Summary of subgraph distributions for 12-atom molecules with a selection of topologies.

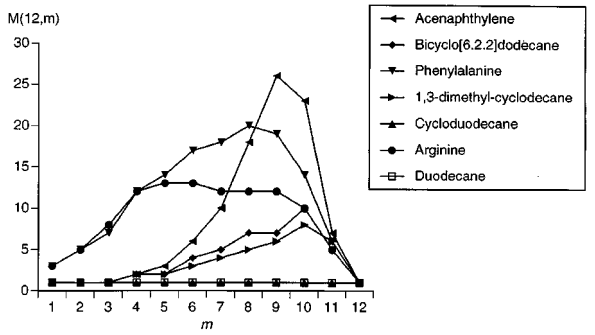

FIGURE 11. Summary of moiety distributions for 12-atom molecules with a selection of topologies.

These observations may also be represented by the averaged substructure ''degeneracies.'' For each fragment size, the number of subgraphs is divided by the number of distinct moieties, to produce the average ''degeneracy.'' In each case this will be a number greater than or equal to 1.0. The closer to 1.0 that it is, the greater the real ''diversity'' within a given set of fragments, because it means that most of the moieties of that size occur a small number of times. In the limiting case, in which the average degeneracy is exactly equal to the number of subgraphs, it means that there is only one distinct moiety. Table X presents this information for a selection of molecules including those whose fragment distributions are displayed in Figures 10 and 11, and a selection of other molecules. It is clear that, while the degeneracies generally decrease as fragment size increases, the amino acids offer the greatest diversity per fragment across the range of fragment sizes, both because of branching and the presence of several different types of heteroatoms.

Discussion and Conclusions

Commonly occurring molecular skeletons have been investigated by exhaustive enumeration of the numbers of constituent subgraphs and of the

TABLE X.
Average Degeneracies for a Selection of 12-Atom Molecules.

	m											
	1	2	3	4	5	6	7	8	9	10	11	12
Chain												
Duodecane	12.0	11.0	10.0	9.0	8.0	7.0	6.0	5.0	4.0	3.0	2.0	1.0
Amino acid												
Arginine	4.0	2.2	1.6	1.2	1.1	1.0	1.0	1.0	1.0	1.0	1.0	1.0
Phenylanaline	4.0	2.4	2.1	1.6	1.7	1.5	1.7	1.7	1.7	1.6	1.3	1.0
Simple ring												
Cycloduodecane	12.0	12.0	12.0	12.0	12.0	12.0	12.0	12.0	12.0	12.0	12.0	1.0
Methyl-substituted rings												
1-Methylcycloundecane	12.0	12.0	13.0	7.0	7.5	5.3	5.7	4.5	4.8	4.0	1.8	1.0
1,3-Dimethylcyclodecane	12.0	12.0	14.0	8.0	9.5	7.3	6.2	5.6	5.2	3.1	1.7	1.0
Bicyclic molecules												
[6.2.2]Dodecane	12.0	13.0	16.0	10.5	14.0	8.5	8.0	6.6	7.4	4.1	2.4	1.0
Fused rings												
Acenaphthylene	12.0	14.0	20.0	16.5	18.3	13.5	10.5	6.1	3.3	2.2	1.7	1.0
Cage molecules												
Icosahedrane	12.0	30.0	40.0	70.0	82.0	62.5	70.9	49.5	44.0	22.0	12.0	1.0
Substituted cages												
1,-Carborane	6.0	15.0	16.0	21.0	20.5	14.2	12.0	9.5	7.9	5.5	3.0	1.0
1,2-Dicarborane	6.0	10.0	11.4	12.4	11.7	7.7	5.9	4.4	3.7	2.9	2.4	1.0
1,7-Dicarborane	6.0	15.0	16.0	19.1	20.5	16.6	16.3	13.4	13.8	8.2	6.0	1.0

distinct moieties that can be obtained for each possible size. This extensive study has only been possible with the development of a new computer program,[21] which handles the perception of general fragment topology to arbitrary size and includes the ability to recognize chemical equivalence among substructures.

The present analysis has proceeded on two levels. Generation and enumeration of the connected subgraphs of size m contained in a molecular graph of size N give rise to a distribution function, $S(N, m)$, whose form depends on the topology of the graph. Specifically, the number of subgraphs underlying a given molecular topology is very sensitive to the connectivity of the atoms. While high average connectivity, as is found in cages and fused ring systems, leads to large numbers of subgraphs, the presence of merely a few high-connected atoms can also increase the number of subgraphs relative to a skeleton with the same average connectivity. Systematic comparison of the subgraph to ascertain chemical equivalence gives rise to the set of distinct moieties that may be extracted from the molecular skeleton. The distribution function for the moieties, $M(N, m)$, is sensitive to the numbers of distinct atoms which are present as well as to symmetry of the skeleton.

It has been shown that claims and simple rings give the simplest and flattest distributions of fragments with approximately linear dependence of $S(N, m)$ on N and m. The introduction of branching into these forms increases the numbers of subgraphs somewhat and alters the form of $S(N, m)$, but cagelike molecules and fused ring systems give the greatest numbers of subgraphs for any size of molecule. Substitution of heteroatoms on each of these families of skeleton gives rise to forms for $M(N, m)$ which are broadly similar to the respective forms for $S(N, m)$ but are flatter. Increasing symmetry[25] of the skeleton has a corresponding diminishing effect on the numbers of distinct moieties. Naturally occurring molecules such as amino acids are able to give rise to large numbers of distinct moieties for their size, both because they have branched skeletons and because they contain several different atom types (i.e., they have multiple heteroatom substitutions).

There are important consequences of these results for the design of the scaffolds which underpin combinatorial chemistry methods. A desirable

property of a chemical library is that it should contain high diversity. In a very crude way, this may be simply measured by a count of the numbers of distinct moieties present. Such a count measures the basic composition of the molecular constituents at the 2D level, from which it can be extrapolated to the 3D level that the likelihood of matching a given pharmacophore also depends on the number of distinct substructures present. The higher the connectivity of the atoms in the basic skeleton, the wider the spectrum of fragments covered. Increasing the number of different types of heteroatom substituents also leads to considerably enhanced substructural diversity.

Using this understanding of the relationship between substructure counts and topology, it is possible to ''design'' molecules with useful fragment distributions. It may be important to maximize the occurrence of substructures of ''medium'' size (i.e., ~ 8–12 atoms) because real fragments of such size are able to span the typical distances between receptor-interaction points; smaller fragments are unable to attain the geometries necessary to achieve an interaction and larger ones might become unwieldly. Substructure counting has provided numerical evidence that branched acyclic molecules and cagelike and fused-ring scaffolds are useful because they compactly give rise to maximal substructural representation. Branched acyclic molecules give fewer substructures than either and have the disadvantage of high conformational flexibility. One example of this general principle has been provided by Carell et al.,[28–30] who used rigid cages and fused-ring molecules as core structures to which functionalities have been attached combinatorially.

One reason why exhaustive substructure enumeration has not been directly addressed before is because it has been perceived to be a problem which can rapidly explode. In fact, for the modest-sized molecules considered here, almost all enumerations could be carried out in a matter of a few seconds of computer time and the numbers of fragments are not prohibitive. The chemistry of carbon is responsible. High connectivity of (merely some, not all) atoms in close proximity in the skeleton leads to generation of large numbers of fragments. This result shows that, in molecules in which the connectivity rarely exceeds four, there will actually be only a modest number of possible substructures. For fixed molecular size, increase in average atom connectivity leads to a dramatic increase in the number of subgraphs possible. Bringing highly connected atoms close to one another increases this effect further; the limiting case is a cage in which all atoms have high connectivity. For fixed N and average connectivity, increased branching will increase the number of subgraphs of given size.

In any case, the maximum number of induced subgraphs[18] of size m obtainable for molecules with N atoms is simply the binomial coefficient, $\binom{N}{m}$. Only in the case of so-called ''complete graphs''[17] (ones in which every vertex shares an edge with every other) is this possible. In chemistry, molecules whose 2D representation would be complete graphs with more than five vertices are all but impossible. Cyclopropane is the smallest complete graph examined here; its subgraph counts for sizes $m = 1$, 2, and 3 are 3, 3, and 1, respectively, corresponding to the binomial coefficients for $N = 3$. Prismane is also a complete graph and, as seen in Table I, the numbers of subgraphs follow the binomial coefficients for $N = 4$, peaking at $m = 2$, with $\binom{4}{2} = 6$. For the larger molecules the numbers of subgraphs of given size fall short of the comparable binomial coefficients. Even for cubane, where $N = 8$, the number of fragments of size 5 is only 48—much less than the comparable binomial coefficient, 70. For dodecahedrane, the number of subgraphs of size 10 is less than 7% of the theoretical maximum for a complete graph with 20 vertices, 184,756. For ''Buckminster Fullerene,'' the number of fragments of size 11 is dwarfed by the number that would be possible for a complete graph with 60 vertices; that is, $\binom{60}{11} = 3.427 \times 10^{11}$. It also follows that distributions of subgraph counts which are exactly symmetrical about their midpoint, $m = N/2 + 1$, are only possible for the ''complete graphs.'' Therefore, the numbers of subgraphs that must be dealt with is actually much more manageable than the $N!$ problem which plagues other combinatorial aspects of chemical enumeration, such as isomer counting.[31]

An important feature of systematic enumeration is that it is not limited to specific types of substructures. The enumeration of a somewhat restricted set of substructures is implicit in the computation of many topological indices.[32–34] For example, there are indices based on the counting of ''self-avoiding paths''[35] and also the count of chains, cycles, and small ''clusters.''[36] Although restricted counts can be useful in certain applications, such as structure–activity relationships,[37] topological indices are typically based on substructures up to a size of seven atoms.[36] The fused-ring

systems studied here have shown that discrimination may depend on counting larger fragments in many cases. The mathematical relationship between substructure counts and structures that exists for many of the topologies considered gives grounds for believing that substructure counts may also usefully be the basis of structure–activity relationships.

A different perspective on the problem of substructure enumeration has been present in chemical database design.[23,38] For instance, some current software represents molecules as lists of ''important'' fragments, which are flagged as present or absent, to facilitate the search for a given substructure. But it is rare to be able to represent explicitly just how many times even these fragments may appear. The enumeration of moieties shows that, in some cases, a knowledge of explicit counts can be valuable.

Nevertheless, there are many limitations to considering substructures at the 2-D level. In particular it has not been possible to incorporate stereochemical features. For example, cis and trans double bonds have not been distinguished, nor have the alternative configurations at asymmetric centers been taken into account. A related issue is that the underlying conformational flexibility of a given moiety cannot be treated with a 2D-based appoach. Therefore, the actual number of ''distinct'' 3D spatial arrangements which are possible from a given collection of molecules may not relate simply to their substructure distributions. Despite these limitations, this study has provided valuable insights into the influence of the topology in the generation of substructural diversity.

The work presented here should be regarded as a preliminary investigation of a complex problem. Some interesting results have been obtained and further applications of substructure enumerations can be envisaged. Subsequent work will investigate the enumeration of the distinct moieties from collections of molecules in a series or database (a matter which was touched upon briefly during consideration of the amino acids). As an objective method that can be uniformly applied to essentially all the scaffolds of interest in combinatorial chemistry, the calculation of total and unique fragment counts could provide indices for ''molecular diversity.'' Such analyses and the comparison of fragment distributions from different sets of molecules with related behaviors may lead to the deduction of the statistically important fragments, in the spirit of the work carried out in areas of toxicology[5,7] and environmental hazards. In essence, this approach may be developed into a means to identify moieties whose abundance suggests importance of function and which might be responsible for a given activity.

Acknowledgments

We express our gratitude to Dr. Lawrence M. Kauvar for many constructive criticisms of the manuscript. We are grateful to Prof. Harel Weinstein for some pertinent comments. We thank Tran Boi Trán for deriving and drawing the 75 decane skeletons.

References

1. N. A. B. Gray, *Computer-Assisted Structure Elucidation*, Wiley-Interscience, New York, 1986.
2. E. J. Corey and W. T. Wipke, *Science*, **166**, 178 (1969).
3. A. Panaye, J.-P. Doucet, and B. T. Fan, *J. Chem. Inf. Comp. Sci.*, **33**, 258 (1993).
4. G. Klopman, *J. Am. Chem. Soc.*, **106**, 7315 (1984).
5. G. Klopman, J.-Y. Li, S. Wang, and M. Dimayuga, *J. Chem. Inf. Comp. Sci.*, **34**, 752 (1994).
6. J. Gasteiger, W. Hanebeck, and K.-P. Schulz, *J. Chem. Inf. Comp. Sci.*, **32**, 264 (1992).
7. G. Klopman and M. I. Dimayuga, *J. Comp.-Aided Molec. Des.*, **4**, 117 (1990).
8. D. F. V. Lewis, In *Reviews in Computational Chemistry III*, K. B. Lipkowitz and D. B. Boyd, Eds., VCH Publishers, New York, 1992.
9. P. Willett, *J. Chem. Inf. Comp. Sci.*, **23**, 22 (1983).
10. M. A. Gallop, R. W. Barrett, W. J. Dowler, S. P. A. Fodor, and E. M. Gordon, *J. Med. Chem.*, **37**, 1233 (1994).
11. M. A. Gallop, R. W. Barrett, W. J. Dowler, S. P. A. Fodor, and E. M. Gordon, *J. Med. Chem.*, **37**, 1385 (1994).
12. M. H. Lyttle, *Drug Dev. Res.*, **35**, 230 (1995).
13. E. J. Martin, J. M. Blaney, M. A. Siani, D. C. Spellmeyer, A. K. Wong, and W. H. Moos, *J. Med. Chem.*, **38**, 1431 (1995).
14. L. M. Kauvar, D. L. Higgins, H. O. Villar, J. R. Sportsman, A. Engqvist-Goldstein, R. Bukar, K. E. Bauer, H. Dilley, and D. M. Rocke, *Chem. Biol.*, **2**, 107 (1995).
15. R. N. Zuckerman, *Curr. Opin. Struct. Biol.*, **3**, 580 (1993).
16. B. A. Bunin and J. A. Ellman, *J. Am. Chem. Soc.*, **114**, 10997 (1992).
17. F. Harary, *Graph Theory*, Addison-Wesley, Reading, MA (1972).
18. This means that the subgraphs considered are of a specific type known as ''induced subgraphs.[17]
19. J. R. Ullmann, *J. Assoc. Comput. Mach.*, **23**, 31 (1976).
20. J. M. Barnard, *J. Chem. Inf. Comp. Sci.*, **33**, 532 (1993).
21. R. G. A. Bone, *RAPTOR*, Terrapin Technologies (1995).

22. A. Dalby, J. G. Nourse, W. D. Hounshell, A. K. I. Gushurst, D. L. Grier, B. A. Leland, and J. Laufer, *J. Chem. Inf. Comp. Sci.*, **32**, 244 (1992).
23. ISIS/Base 1.2.1, Molecular Design Limited, San Leandro, CA (1995).
24. G. M. Downs, V. J. Gillet, J. D. Holliday, and M. F. Lynch, *J. Chem. Inf. Comp. Sci.*, **29**, 172 (1989).
25. Geometric (3D) symmetry has no explicit representation in a 2D molecular connection table, and a graph which depicts a highly symmetric molecular structure may be deformed in any way to conceal that symmetry without altering the overall connectivities. Nevertheless, where geometric symmetry dictates the presence of nodes with equivalent topological environments, symmetry becomes manifested as automorphism operations on the adjacency matrix.[17,26] In this way, the symmetry group of the graph is the automorphism group. There is not necessarily a one-to-one correspondence between elements of the automorphism group and elements of a point group.
26. M. I. Davis and M. L. Ellzey, Jr., *J. Comput. Chem.*, **4**, 267 (1983).
27. H. O. Villar and L. M. Kauvar, *FEBS Lett.*, **349**, 125 (1994).
28. T. Carell, E. A. Wintner, A. Bashir-Hashemi, and J. Rebek, Jr., *Angew. Chem. Int. Ed. Engl.*, **33**, 2059 (1994).
29. T. Carell, E. A. Wintner, and J. Rebek, Jr., *Angew. Chem. Ed. Engl.*, **33**, 2061 (1994).
30. T. Carell, E. A. Wintner, A. J. Sutherland, J. Rebek, Y. M. Dunayevskiy, and P. Vouros, *Chem. Biol.*, **2**, 171 (1995).
31. G. Pólya, *Acta Math.*, **68**, 145 (1937).
32. M. Randic, *J. Am. Chem. Soc.*, **97**, 6609 (1975).
33. H. Hosoya, *Bull. Chem. Soc. Jpn.*, **44**, 2332 (1971).
34. H. P. Schultz, *J. Chem. Inf. Comp. Sci.*, **20**, 227 (1989).
35. M. Randic, *J. Chem. Inf. Comp. Sci.*, **24**, 164 (1984).
36. L. B. Keir and L. H. Hall, *Molecular Connectivity in Structure–Activity Analysis*, Research Studies Press Ltd., John Wiley & Sons, New York, 1986.
37. S. C. Basak, In *Practical Applications of Quantitative Structure–Activity Relationships (QSAR) in Environmental Chemistry and Toxicology*, W. Karcher and J. Devillers, Eds., 1990, ECSC, Brussels & Luxembourg, p. 83.
38. C. A. James and D. Weininger, *Daylight Theory Manual*, Daylight Software 4.41, Daylight Chemical Information Systems, Inc. (1995), Irvine, California.